Decentralized Networks vs The Trolls

Derek Caelin

Abstract

In the summer of 2019, the far right social network Gab migrated to the decentralized "Fediverse" of social networks after being booted from mainstream financial services and hosting solutions. Almost immediately, Gab was met by a dedicated movement to isolate it. The movement was largely successful; within a year, the Gab CTO announced they would leave the Fediverse. This paper recounts how moderators, activists, and developers in the Fediverse used strong moderation tools, representative codes of conduct, and no small amount of organization to promote healthy online spaces. However, the Fediverse is no utopia, and many moderators still struggle to address motivated antagonists. Ultimately, decentralization provides both opportunities and challenges for users of this form of social media.

Technology is exploding the concept of community, as it always has.

The history of society is one of continuous expansion and interconnection, and technology has often been the driver. When humans first started creating communities, group identity used to rely primarily on factors such as birth and physical location (Gupta 2020); gradually the advent of newspapers, railways, and eventually mass telecommunications technologies have expanded and mixed human relations so that today a person might more readily identify with a contact in an online social network than with his or her neighbor.

As the world grows more connected and human interactions increasingly take place in online spaces, the future for humanity depends on our ability to govern online social spaces effectively. Our ability to wield this new technology effectively matters, especially to minority or historically marginalized populations. Says Zeynep Tufekci: "Technologies alter our ability to preserve and circulate ideas and stories, the ways in which we connect and converse, the people with whom we can interact, the things that we can see, and the structures of power that oversee the means of contact" (Tufekci 2017), essentially, the things that matter most. The rules and tools we use to govern our online spaces shape the experiences we have within them. This chapter explores how one decentralized social network—the Fediverse—has sought to navigate a new technology, better the experience of its community, and deal with the issue of bad actors who have sought to establish a platform for dangerous content.

The following is the result of research and interviews I conducted over the course of several months in early 2020. In that time I spoke with 80 people, primarily on the decentralized microblogging social network Mastodon, including 12 interviews of social network administrators and moderators as well as hate speech monitors and experts in dangerous speech. I also conducted a survey of Mastodon users, with 670 respondents.

The Challenge of Centralized Social Media

Big social networks are in a tough spot. Amid quirky memes, family stories, and news updates, big social platforms, particularly Facebook and Twitter, are also saturated with content that can increase the risk that its audience will condone or commit violence against another group. The effect of social media upon the spread and impact of "dangerous speech"—that is, "any form of expression (e.g. speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group" (Dangerous Speech Project 2020) has been dramatic. Speakers with fringe ideologies can find an audience and interact in "echo chambers," bolstering one another and spreading their narratives even further.

Despite modifying their policies to disallow hate speech and evicting many prominent white nationalists (European Commission 2020), Twitter, Facebook and YouTube have failed to excise Nazi and white nationalist content from their sites. The question remains what to do about it. In some cases, social media companies have gradually taken steps to regulate bad behavior.

Facebook has banned hate speech ("Community Standards," n.d.), and haltingly extended their content moderation policies to prevent explicit support and representation of white nationalism and white separatism (Cox and Kobler 2019). Calling for or encouraging the harassment of others (Murphy and Cacace 2020) is also prohibited. Twitter, likewise, has made cautious evolutions over time; it first banned hate speech ("Hateful Conduct Policy," n.d.), then updated the policy to include religious dehumanization (Wright 2019).

Notwithstanding this, hate speech, harassment, and incitement to violence still appear on the big social media sites, often with dramatic consequences. When in 2019 a gunman live-streamed his attack on the Al Noor Mosque in Christchurch, New Zealand, the video was viewed thousands of times, despite Facebook's active effort to expunge it (Facebook 2019).

A June 2020 Civil Rights Audit of Facebook criticized the company's track record in addressing bad content, even as it evolves on a policy level:

> *While Facebook's Community Standards prohibit hate speech, harassment, and attempts to incite violence through the platform, civil rights advocates contend that not only do Facebook's policies not go far enough in capturing hateful and harmful content, they also assert that Facebook unevenly enforces or fails to enforce its own policies against prohibited content. Thus harmful content is left on the platform for too long. (Murphy and Cacace 2020)*

There is plenty of room to critique social media companies for their slowness to address the issues of dangerous speech and disinformation on their platforms. At the same time, many of the problems they face in addressing content are structural, and not solvable simply through policy. Twitter measures its monthly active users in the hundreds of millions (Clement 2019Statista Research Department 2021), Facebook, in the billions (ClementStatista Research Department

20201). At the time of writing, 28% of the world population is on Facebook (Internet World Stats 2020). The task of reviewing content is gargantuan and outstrips many of the tools that the big tech companies have at their disposal.
Tech companies rely both on human and machine resources to process content, and both are imperfect resources. Facebook users, for example, flag over one million pieces of content worldwide, but "the vast majority of items flagged do not violate the Community Standards of Facebook. Instead content flags often reflect internal group conflicts or disagreements of opinion" (Klonic 2017). Artificial Intelligence is capable of sifting through vast swaths of content, but still frequently fails to account for context or misidentifies the nature of posts (Copia Institute 2020).

Mike Masnick, a technology researcher and the editor of Tech Dirt says that "Content moderation at scale is impossible to do well" (Masnick 2019a) First, Moderation is likely to anger whomever is moderated—and the more people in a social network, the more types of differing views, mindsets, and perspectives the policy will need to be represented by policy.

Second, says Masnick, content moderation is inherently subjective, relying on judgment calls in gray areas that are bound to be questioned. Third, there is an issue of scale:

> *Getting 99.9% of content moderation decisions at an "acceptable" level probably works fine for situations when you're dealing with 1,000 moderation decisions per day, but large platforms are dealing with way more than that... On Facebook alone a recent report noted that there are 350 million photos uploaded every single day. And that's just photos. If there's a 99.9% accuracy rate, it's still going to make "mistakes" on 350,000 images. Every. Single. Day.*

Large social networks appear to have trouble responding swiftly to hateful or violent content. I asked people who monitor hate speech how rapidly social networks responded to the content they found and reported. Jacqueline LaCroix, who researched hate speech at PeaceTech Lab and other places ("PeaceTech Lab | Hate Speech," n.d.), tells me that "it generally took several weeks for Twitter to remove them, if they did."

> *I reported accounts that were either blatantly equating racial or ethnic groups with animals or other inhuman characteristics or were advocating violence against certain groups… most of these were eventually banned from Twitter but I think it took around 2-4 weeks at least.*

One technologist based in Pakistan who conducts social media monitoring campaigns reflected that: "We have dismantled many cyber armies on Twitter. We would report material against minorities, violent extremism, dangerous speech. At times Twitter responded. At times they didn't."

And so, the behemoth social media companies stumble on, removing millions of posts per day, but unable to resolve the issue of harmful content on their sites.

The Fediverse

In May 2008, a Montreal-based developer named Evan Prodromou published the first post on a humble new social network. Identi.ca became a darling of many free and open software advocates (Behrenshausen 2013), and presented a creative take on social media allowing many different machines to run the network. Rather than relying on a single company to host the service, any developer could spin up an Identi.ca server and start communicating with other users. This sort of relationship relied on a protocol—a communication standard that anyone could use to build a compatible service—which would go through several iterations, eventually becoming the ActivityPub protocol in 2018. After it was recommended by the World Wide Web Consortium as a way for two social services to communicate with each other (W3C 2018), the technology has blossomed and been adopted as a communication standard by a range of different platforms.

ActivityPub drives a suite of alternatives to the well-known and familiar services that make up the daily Internet. PeerTube is a free and decentralized alternative to video hosting platforms like YouTube and Vimeo. Pixelfed stands in for image-based social networks like Instagram. WriteFreely replicates Medium. The largest and most active of the ActivityPub-based services is Mastodon, a Twitter-like social network with 500-character posts presented in a reverse-chronological timeline. Together, these and other services form a federated universe of social networks. Each of these applications employs the ActivityPub protocol to communicate, so that different services can speak to one another and exchange content; for example, a user of Mastodon could subscribe to a PeerTube video channel and comment on videos as they are posted. This network of connected servers and services sharing content and users is called the "Fediverse" (an amalgamation of "Federated" and "Universe").

Today, with 4.4 million users, Mastodon is the largest and most established of these networks, and it is made up of thousands of servers (called "instances") that focus on different topics and communities. Mastodon.social is one of the oldest and largest instances, but it has deliberately stopped accepting new applicants so as not to grow too large and drown out the other Mastodon servers. There are instances for a variety of interests: Mastodon.technology is a place for techies to talk shop (or anything else that comes to mind)—Social.coop is where enthusiasts for cooperatives congregate. There are instances for political communities, instances for advocacy, for LGBTQ communities, and for writers.

Soon after its launch, Mastodon gained a reputation for being a sort of anti-Twitter. The creator of the technology, Eugen Rochko, explicitly critiqued Twitter's failures to address harassment and hate speech in his reasons for starting the service (Rochko 2017). Early coverage of the network praised its "diverseyet welcoming online environment" (Rhodes 2017) where users are "authentic" and active in creating safe spaces (Jackson 2018). Even as Mastodon won praise from its users, commentators speculated what would happen when a decentralized network grew beyond the control and moderation reach of its creator (Morse 2017).

Here Come the Trolls

In June of 2019, a worried developer began a conversation about white supremacists. Github user Laurelai voiced a concern that had saturated her community for weeks, ever since Gab, a far-right social network known as a safe haven for neo-Nazis and white supremacists, had announced its intention to launch a new platform. Gab's previous site had been taken offline in October 2018, after a platform user killed eleven people in a Pittsburgh synagogue, and scrutiny of the Gab by its hosting and financial service providers shone a spotlight on its acceptance and encouragement of hate speech (Rosenberg 2018).

Gab was descending on the Fediverse, whose users were particularly dismayed that the open source code Gab's developers were exploiting originated from Mastodon. Developers worried that unless something were done, Gab's hundreds of thousands of users would spill over and contaminate their own social networks.

"Not to put too fine of a point on it but we have about a month before the monsters come," said Laurelai, pointing to the July start date of the new server (Laurelai 2019).

> *When i first started getting into mastodon i tried to tell people these types would show up eventually and got laughed at and ridiculed for it. Now the barbarians are at the gates and we are not ready and a lot of people just want to let them in…If we aren't prepared these fascists will overrun and destroy our communities.*

The migration of Gab marked a major test for the Fediverse. Because no one authority controlled the policies of the network, it would be impossible to bar Gab with a single sweeping action, the way centralized social networks such as Facebook or Twitter might. Instead, it would take a mass movement of administrators, moderators, and technology providers to block, isolate, and moderate the new server. For years, users of the Fediverse had touted that its design favored kinder interactions and provided more tools for moderators to address the threat of bad actors online. That design would now be put to the test.

Laurelai's fears were soon realized. Within a day of launching, Gab had swollen to become the largest node in the Fediverse, reportedly more than doubling the size of the next largest server (Robertson 2019). The worst had happened; the trolls had deposited themselves in the new frontier.

In the year following the advent of Gab, however, the Fediverse demonstrated why its design is resilient to bad actors, and how decentralization and powerful moderation tools have succeeded in making a social network that better—though not perfectly—protects its users, not only from neo-Nazis, but to some extent from racismand trollish behaviors. The Fediverse's success in dealing with a wealth of bad actors like Gab demonstrates how the decentralized social network

is better-equipped to provide a safer, more positive online experience than the centralized alternatives.

In interviews, conversations, and surveys of users of the Fediverse, users and administrators pointed to three elements that helped them to address the issues of bad content and bad actors: codes of conduct at the instance level, moderation by humans, and an array of effective moderation tools.

Codes of Conduct

Codes of conduct are rules of behavior that govern a particular instance within the Fediverse. By being a member of an instance, a user is required to adhere to that instance's rule set. In a decentralized network, this means that many different instances can maintain different codes of conduct, and that users will interact with people governed by different rules.

Administrators in the Fediverse will often use words like "evolving," or "changing" to describe their codes of conduct. Depending on the instance these rule sets can be either far more exacting or loose, depending on the interests of the administrator or the communities they support. Within the Fediverse the spectrum of policies ranges from promoting unrestricted speech to defining explicit rules for acceptable behavior.

For example, Qoto.org, an instance for "scholars in Science, Technology, Engineering, and Mathematics (STEM)," enacts few regulations on member behavior. "We are a free speech, no censorship zone," says the instance's "about" page (Qoto.org, n.d.). "Feel free to talk about whatever you want, say whatever you want, as long as it is legal and you engage respectfully you won't ever get banned from the server. With that said we do have a few rules: No spam, and no using multiple accounts to circumvent personal bans, and a few others."

On the other end of the spectrum is an instance like Mastodon.technology, which maintains a long list of rules (Mastodon.technology, n.d.). The policies are intentionally broad (in a way that might give the policy team at Facebook heartburn). "No racism. No sexism," says the rule list. "No discrimination based on gender, sexual minority, sexual orientation, disability, physical appearance, body size, race, ethnicity, religion (or lack thereof), or national origin." Seventeen bullets down, just in case the reader thought there was some form of vice that would be tolerated, a catch-all, somewhat tongue-in-cheek rule is added: "No other unethical conduct."

When selecting a home in the Fediverse, a user might end up selecting an instance not only based off the stated interests of the server, but also on the governing code of conduct and how well it aligns with the rules they want to be governed by. People who favor unregulated environments gravitate in one direction, others who want to exist in places with specific rules governing members go in another. In this way, users get to choose what sort of experience they want to have. The Fediverse achieves a flexibility that can't happen in centralized social networks. Where big social media platforms tend to restrain behavior only in limited circumstances, many Fediverse servers adopt sweeping regulations, depending on the wishes of the administrator and the denizens of the instance.

While social media companies tend to update their policies only after public scrutiny and pressure, leaving harmful content online, codes of conduct on Mastodon appear to change quickly to meet the needs of their community to address user needs.

Emi Do, a moderator on Social.coop, describes a process by which people highlight concerns, discuss them, and vote on them using an open source, online decision-making tool called Loomio. “If we feel like we need to actually write something new into our code of conduct, then we raise that in the community working group in Loomio, and then we raise that with the all members in the general Loomio group.”
Maff, the administrator of Queer.party, tells me that their server’s code of conduct “started out as more or less a boilerplate ‘please don’t do crimes, common sense is an important life skill’, but as more users signed up and more waves of people shifted to mastodon, a lot of stuff came up that it felt important to actually -state-,” such as their policy around pornographic images.
Mastodon.technology’s code of conduct has also changed dramatically over time. “Originally it was a sort of a plug and play,” Furrow said. He employed a “…Creative Commons license code of conduct, and then when we had the blow up [of new users] we went through it… People gave feedback, and we debated different points, and we took things out and put things in. Yeah, it was a lot of discussion. It was about a three month process. That was really, really interesting and raised all sorts of issues that I hadn’t even considered.”

Accommodating New Users

In November 2019, Mastodon administrators began seeing a spike in new users from across India. Twitter had come under intense criticism in recent months for twice suspending a prominent lawyer who argued before the Indian Supreme Court (Khan 2019). Sanjay Hegde’s account had run afoul of Twitter’s automated moderation tools, which failed to understand the meaning behind his use of a picture of August Landmesser, a German national who refused to salute Adolf Hitler (Basu 2019). Indian users were also critical of how Twitter distributed “blue-checks,” the icon after a person’s username to indicate a verified account. “Blue check marks on @Twitter have become… a matter of caste divide, assertion & privilege,” said Supreme Court lawyer Nitin Meshram (News18 2019). After Hegde publicly migrated to Mastodon, frustrated by Twitter’s intransigence and lack of responsiveness, many Indians began to follow suit.

The new and distinct user group wanted their Mastodon moderators to behave differently from Twitter, and they campaigned for changes in their instances’ rules governing behavior. “There was a big push to remove casteism from the Fediverse, which is discrimination based on caste,” said Ash Furrow, the administrator of Mastodon.technology. “It’s not something that I had ever considered for, it’s something that we added to our code of conduct because it’s something our new users need.”

Human Moderation

Unwelcome content will inevitably emerge in social environments, especially when participants are interacting with communities with different codes of conduct. When this happens, the presence of a moderator whose role is to process and address conflict can be essential.

"This is not Twitter," says the About page of Queer.party. "Reporting posts that make you feel uncomfortable, or that you feel should be brought to our attention, is highly encouraged. Your reports will not be ignored, and will be dealt with (usually) within a few days."

Users appreciate having active moderation to address issues when they arise. "Places that have very active moderation… even places that are larger but have a more involved mod team like Playviscious.social will have a quick and often more satisfying response to reports," one user of the latter instance told me.
"It's also a question of content," the user continued, highlighting how differing codes of conduct can result in clashes between two instances. "If the instance has a [code of conduct] that allows for racists to participate in their community, they're obv not going to be as sensitive to those types of abuse reports. If they host a bunch of irony poisoned lefty edgelords, they're not going to respond to reports that request sensitive content warnings and such."

Because of their massive scale, centralized social networks face significant challenges in moderating content, even with the aid of automated moderation. Even its vast array of in-house and third party moderators (Wong 2019) have not allowed Facebook to fully address its content moderation problems. By contrast, the Fediverse moderators I spoke with in this research project tended not to feel overwhelmed by their tasks. Mastodon.technology, an instance with 20,000 users (roughly 2,000 of which are active every month) needs its four moderators to take action only about a dozen times a week.

"A single mod can handle most of the load," said Daniel Supernault, the developer for PixelFed and the sole moderator for the flagship instance for the ActivityPub equivalent of Instagram, "We don't get many reports, the latest one on Pixelfed.Social is three months old." The moderators of Social.coop, an instance of about 2,000 people, only need to intervene once or twice a week. LGBTQIA.is acts a couple of times a month.

"I pretty much never see things that are outright hateful due to a combination of network effects and my mods being really proactive about trying to firewall off the kind of spaces that cultivate dickwads," one user, Starkatt, told me (katt, fracturing 2020). "The great thing about the mods is they're community members so they're embedded in the relevant social context. This means they're able to act based on subtleties that might be invisible to outsiders. It also means their curation is very active because they don't want to deal with dickwads either."

Moderation Tools

What about when moderators need to take action? What levers do moderators pull when dealing with content on their instances? Moderators and administrators have a range of options at their

disposal, ranging from facilitated discussion to an instance-wide ban. Interestingly, not all moderation tools are the same across platforms, leading to different challenges between different social networks.

Discussion

The beginning of most interventions can be as simple as asking people to engage. Emi Do, a moderator for Social.coop, explained how she and other moderators on her team try to deal with issues when conflict arises on her server.

> *Most of the things that we are intervening on are people who are concerned about the behavior of people off our instance, not necessarily on our instance. There's a lot of agreement in terms of our community code of conduct. If there is an issue, if we just point to the code of conduct and facilitate a conversation between the person who has raised the issue and the person who perpetrated the issue, people are very open to taking down posts, editing, engaging in conversation. But I think that's just who we are attracting to our instance. I think the reason why people are on our instance might be different from the reasons why they would be on Twitter.*

Silencing

Moderators also have the ability to "silence" a user or an instance, meaning that their content will not be seen in timelines unless they are followed by a user. This option allows users to follow users and interact with if they seek it out, but it will not make other users on the view the same content.

Silencing can be an appropriate tactic if there is a single bad user on a server, but if the bad actor is representative of the server as a whole, it may be worthwhile to take extreme measures: defederation.

Defederation

Moderators and administrators have real power inusing Mastodon, within their own domains. It's within the ability of moderators to block individuals from interacting with their users or—if necessary—defederate with the server altogether, preventing any content or users from one site showing up on their server.

The administrator of Queer.party, who goes by "Maff" online, explained the more dramatic measures instance admins can take to protect their users from bad actors. "When an instance values free speech above all, to the detriment of all (e.g. Gab) the standard policy is to de-federate it - mark the domain as 'suspended', preventing any communication between it and [another instance] This also applies when the instance either outright does not perform any moderation, or when their policies explicitly allow harmful activity or bad actors. It doesn't feel

great, but the alternative is that the federated timeline gets filled with hate, or users … get harassed by people from those instances."

The hashtag "#Fediblock" is one way users and administrators make reports on bad actors, explaining which instances or users they'd recommend blocking proactively, and why. When I interviewed users about their Fediverse experience, they explained that their moderators proactively watched the #Fediblock hashtag and took action prior to issues emerging on their own instance. Accounts like @IsolateGab@todon.nl maintain a recommended blocklist for admins looking to deny Gab and its affiliates access to the broader Fediverse.

Defederation in Action

In March of 2020, the administrator of Queer.party logged in and discovered that a member of their community had filed a request for moderation. The admin, Maff, was a software security specialist based in Scotland who had moderated the instance since they founded it in 2017. Since that time they had seen moderation reports in all the variety that the Fediverse had to offer, so when they saw a reported comment which read, "Reality: The redskins started fights they couldn't win and paid dearly," they knew they'd need to do some research in order to choose how best to respond to the message.

At first, they assumed that the message might be related to American sports, since they were aware that a team in the U.S. was named "The Redskins." But after exploring the context of the post, they soon realized that the reference was not to sports but to American history, and that the term "redskin" was a derogatory term for Native Americans.

By itself, the comment may not have caught the attention of the major social media companies. But Queer.party's code of conduct explicitly disallowed "racism, homophobia, transphobia, xenophobia and any other discriminatory conduct which attacks any individual or group of individuals based on what they are or what they believe in," and the post—both the derogatory phrase and the callous justification of violence—fell outside the bounds of acceptable conduct within the instance. Given the content of the message, Maff felt that their role as administrator required an intervention. The question remained—how to respond?

The reported post had come from outside Queer.party, so Maff's next step was to check the code of conduct of the user's instance. If they found that the terms of service of the poster's home instance disallowed this sort of posting, they might be able to initiate a dialogue and engage the moderator of the other instance in a dialogue with the poster.

What they found was dismaying. Far from being an ordinary member, the poster in question was on the instance's moderation team. What is more, the server's code of conduct was filled with red flags. "No LGBTQ. Period," it read. "No homosexuality. No men who think they're women or women who think they're men. No made up genders." The instance's rules made it clear that they'd be unsympathetic to Maff raising the issue of the admin's conduct for discussion.
Given the diametrically opposed values of the two instances, Maff made the decision to make an instance-wide block—no messages from any of the remote instance's users could now be seen by members of Queer.party. The instances had been "defederated"—firmly separated from one

another within the broader Fediverse. Now none of the users on their instance would see content from the offending site, and ultimately, it was to be hoped, they would have a better experience on the Fediverse.

This example of moderation, which plays out with countless different permutations in the Fediverse, illustrates the difference between moderation on a federated social network and in a centralized social media company. It is likely that if the post were reported on Twitter or Facebook, it may not have been processed for some time, and also possible that would not have qualified as "hate speech" in the terms that the larger platforms would have recognized it. Maff was highly aware of the space they were engaging in, and was able to delve into the context surrounding the content, both in terms of the subject matter and the background of the poster. After examining options for intervention, which included everything from dialog, to silencing or blocking an individual, the administrator was able to take a step to not only respond to an individual user's report, but to proactively protect the community at large.

This structure is ultimately one of the reasons that the Fediverse feels "friendlier" to many of its users, because instance administrators are able to isolate themselves from bad actors. Instances that make no effort to moderate content can find themselves speaking only to their own communities or like-minded networks. It isn't that the Fediverse doesn't have bad actors—tools like Mastodon simply give communities more tools to isolate and remove them from their communities.

Not All Moderation Tools Are the Same

Not all moderation tools are created equal in the Fediverse. For example, many users independently told me that video management tools like PeerTube often have a hard time maintaining a clean environment for their users. "There [are] no moderation tools [on PeerTube], so on a peertube instance that federates, it takes a lot of work just to avoid nazi/rape/gore video to be appearing," @Kyzh, the administrator of Rage.love told me. (Several PeerTube instance administrators I spoke with have entirely cut themselves off from other servers in order to avoid this issue.) PeerTube admins have been asking for years to access the tools that Mastodon admins enjoy (Chocobozzz/PeerTube, n.d.), such as the ability to block entire instances. PeerTube's developers were aware of this issue, which is why they raised funds to improve software's moderation features (Framasoft 2020).

The experience of PeerTube users, who have sometimes confronted with child pornography or pro-Nazi imagery, is evidence that simply decentralizing social media isn't enough—administrators must be equipped with powerful moderation tools to be able to tend to their communities.

Scale Impacts Moderation Speed and Quality

In a poll I conducted on Mastodon, 42% of the 674 respondents said that they had reported something, whether it had been a spam account or hateful content, to their instance's moderator. For those who reported on response times (about 60 users across 49 instances), the average upper

limit response time for a moderator to address an issue was 21 hours, with some saying they get feedback in an hour, or even in minutes. To test the moderation speed of their server, one user reported me to their own instance admin to see how long it would take to respond—within a minute, the admin had circled back with a joke (puffball 2020) (Fig. 8.1).

A user reported me to an administrator. The report was addressed within a minute.

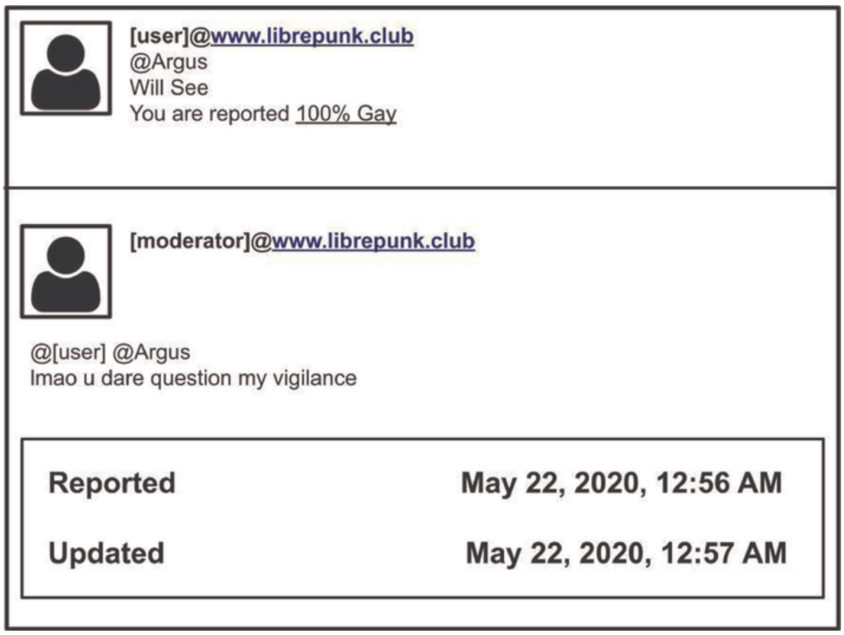

Fig. 8.1

At the same time, scale does seem to play a factor in the quality of moderation. Based on my interviews and discussions with Mastodon users, scale of the instance is a factor that can impact the quality of the moderation response. Mastodon.social, the "flagship" Mastodon instance maintained by the technology's creator, Eugen Rochko, has four moderators maintaining a community of about 39,000 active users (the instance itself has more than half a million users). At a 1:9800 moderator/active user ratio, it's one of the lowest on the Fediverse (outside of Mastodon.social, the average ratio of moderator to active user that I encountered was around 1:150). While some users said that their reports to Mastodon.social had been addressed between 24 and 48 hours, others mentioned that they never received feedback on their moderation requests, and one user from Friend.camp told me that they found the bigger instances to be unresponsive.

"The reports I send to larger, louder instances often see no action," said the user. "But the smaller more mindfully cultivated instances respond almost immediately/within a day or so… Places like Radical.town and Mastodon.social where there are easily 1k active members are basically a black hole for reports." To address this issue, some instances have pre-emptively sought to slow their growth: Octodon.social has switched to member-invites only, "because it's big enough" (Octodon, n.d.). This may be part of the reason the average size for a social network on the

Fediverse is relatively small; according to Fediverse.space, the average size of a Fediverse instance is 870 people.

The Limits of Decentralization

The Fediverse is not inherently immune from the moderation challenges that trouble larger networks. Many users are critical of the design decisions Mastodon has made, particularly around user and content discoverability (Valens 2019), which impact a user's visibility to the broader network, and potential vulnerability to attack. Bad actors can be discouraged from interacting with other servers, but they are free to establish echo chambers in their own spaces. Even the most powerful tools at a moderator's disposal have their limits. Even as some users praise the Fediverse for behaving differently from centralized solutions, others (often from vulnerable or marginalized communities) feel that there is more to be done.

Echo Chambers in Isolation

The Fediverse is far from being all friendly. By design, any instance in the Fediverse can follow its own code of conduct, and if that means having free-wheeling, offensive, even racist conversations, nothing about the design of these platforms prevents it.
There are consequences to fostering spaces that support hateful behavior. The last post of Robert Bowers on Gab before he killed 11 people in a Pittsburgh synagogue in October 2018 was "I can't sit by and watch my people get slaughtered. Screw your optics, I'm going in." Prior to this he had been in a space that glorified hatred, replete with antisemitism and a community that supported violence against Jews, among other ethnic groups. When a poster on Gab asked, "Do you support the Pittsburgh synagogue shooter, Robert Bowers?", nearly 25% of respondents said they did. One poster replied "Robert Bowers literally did nothing wrong" (Conway et al. 2019).

There is a fair critique to be made that an impact of open source, decentralized technology like Mastodon enables bad actors to metastasize on the internet, out of reach of a governing authority. In this respect, the Fediverse has fared no better than centralized platforms in fully expunging bad actors. However, while open protocols and open social networks have allowed instances like Gab to congeal, Mike Masnick points out that "we already have those people infesting the various social networks, and nothing so far has been successful in getting rid of them" (Masnick 2019).

> *...the larger point is that this would likely quarantine them to some extent, as their content would be less likely to get into the most widely used implementations and services on the protocol. That is, while they would be able to be vile in their own dark corners, their ability to infect the rest of the internet and (importantly) to seek out and recruit others would be severely limited.*

Defederation and Filter Bubbles

Critiques of Facebook and Twitter also focus on the fact that these platforms help to generate "filter bubbles"—groups in which users are exposed mostly to like-minded opinions and perspectives. Peacebuilding practitioners at Build Up explain that social media platforms "make it easier for us to surround ourselves with like-minded people, interact with people with differing views in ways that emphasize our differences, and generally deepen political divides" (Build Up 2019). One could argue that moderators defederating their communities from dissimilar instances generates polarization at a community, rather than individual level.

As Siva Vaidhyanathan points out, the platforms we use may or may not limit our vision or entrench our tribalism, and studies of this phenomenon would rely on data kept closely by the platforms themselves. Moreover, the platforms may simply be facilitating the natural human tendency to filter and sort content for themselves. "The arguments about whether the filter bubble exists or how much it matters too often get bogged down in a false dichotomy: is it the fault of technoogy or is it the fault of humans? The answer is always yes." (Vaidhyanathan 2018).

Users of Fediverse instances who are separated from one another on the basis of differing codes of conduct or interests experience the same filtering and sorting as users of other big platforms.

Imperfect Defederation

By definition, a decentralized social network will struggle to apply a single policy across independently-run nodes. In the Fediverse, this means that moderators' actions will have an unexpected or incomplete result outside the domains they control.

When necessary, moderators will block a hostile user or instance in order to prevent their content from showing up on their home instance. The user or instance will fall silent, but moderators will learn that the silencing may not go both ways. Due to confusing sharing agreements between instances—their own users' posts may remain visible on the banned server. Moderators expressed frustration that a by-product of blocking users means they are unable to address misinformation or libel when it happens online. Users occasionally told me about "code wars" where the maintainers of different platforms seek to circumvent each others' content blocking policies. Sometimes these circumventions have less to do with protocols and more to do with human behavior; subversive community members are fully able to take pictures of other users' posts and spread them among their detractors. And sometimes hostile users will be blocked, only return to harass from a different server the next day, forcing admins of targeted communities to play a continuous game of "whack-a-mole."

I asked users whether steps could be taken to address these problems. While some called for structural changes to Fediverse platforms source code, such as requiring Federation to be a consensual process rather than the default, others thought that the worst of the behavior they witnessed couldn't be addressed through changes in code. "I wish more people understood that technology can't solve inherently human problems," one user told me. "Turns out," said another, "no amount of coding will solve the systemic problems of poor human behaviour. Solving these

things needs a cultural shift. And this place simply has not delivered, despite all the BirdSite™ rhetoric and the hubris of people talking about how great this place is."

Putting power in the hands of human moderators and giving them powerful moderation tools and communities bound to representative codes of conduct may address many of the ills of social networks, but the Fediverse is no utopia. Many find the rules of the Fediverse insufficient. Structures can shape human behaviors—encourage some actions, and discourage others, but making a healthy online environment also requires a mass effort.

Gab Is Isolated

The effort to block out Gab (and other bad actors) played out in myriad ways. Activists like Laurelai campaigned hard to get web-clients to block Gab's domain at the software level. Another user, who at the time went by the name Yowlen, pressured App repositories like F-Droid to tag Fediverse client apps with "Promotes bigotry" if they didn't hardcode blocks to Gab and similar repositories (Yowlen 2019). The user, who now goes by the handle KitsuneAlicia, posted the following message on the F-Droid servers:

> *We need a way to combat the bigotry and harassment that goes on in online spaces, and the first step needs to be holding the developers themselves accountable for enabling said bigotry in the first place. If your stance is clearly pro-Nazism, people need to have that warning that your software may not be safe for them.*

As a result of public pressure like this, at least four of the major Fediverse clients (Tusky, Toot,! Mast, and Amaroq) block Gab in some fashion (Robertson 2019). And, at the instance level, a vast number of Fediverse communities simply don't associate with it and other similar actors.'

Catgirl, the administrator for LGBTQIA.is, recalled their experience when Gab entered the scene: "I remember when Gab first came out, I checked it out and immediately you had one of the guys saying to 'shove jews back on a train'…that goes beyond a "freedom of speech" server, it's more of a freedom to incite violence server." When I asked whether Gab was an isolated instance they replied, "No one wants to federate with them. It's a circlejerk of people spreading hate."

Ash Furrow explained his decision to not only defederate his server from Gab, but even from instances that federate with Gab. Setting up two degrees of separation from Gab is a rare step, but Furrow sees it as a necessary step, since through them negative content might seep into the instance he maintains.

> *It's either the case that they agree and they're like 'oh wow, I don't want to be associated with this at all' or they see it in sort of free speech absolutionist terms….And, you know, I can kind of see where they're coming from – and*

listen, I'm Canadian, I'm not really a very vocal person, I don't like confrontation – but I'll tell them, 'is federating with Gab more important than federating with Mastodon.technology?' It's a decision that I put in their field.

Through these dynamics, communities of communities have formed within the Fediverse, the divisions formed through different codes of conducts and different definitions for acceptable speech. Instances like Rage.love, Stardew.city, and Social.linux.pizza, publicize lengthy blocklists detailing the servers they don't federate with, or whose users they silence. Instances with lengthy codes of conduct can end up with weak ties or no ties to instances with relatively few restrictions. "Islands are forming, more or less connected, and that's ok," Arteteco, a moderator on Qoto.org, told me, describing how his instance, which places relatively few restrictions on user behavior, has ended up being blocked by many servers. The effect even more pronounced with Gabis at the center of one of those islands in the Fediverse—disconnected from the broader whole.

Almost one year after it had joined the Fediverse, Gab's CTO Rob Colbert took to the platform to announce that the instance was packing up its bags (5/29/2020). In his post he expressed frustration with Gab's isolation, that "the implementation will change and we won't be compatible with this Fediverse."

Truth be told (and this data is public) Gab is 99.3% insular. 0.7% of people on Gab mention someone who isn't on Gab. So, Gab is wasting resources importing bags of shit no one reads … so that 0.7% of the community can (occasionally) mention someone off-site.

It's unclear what the defederation timeline is for Gab's departure, or what form the site will take. But it is clear that isolation has impacted the network's plans and dimmed their hopes of finding a broader community in a decentralized social network. The Fediverse had rejected Gab.

What's Next for Social Media?

As more public focus is directed on social media and its impact on society, both centralized and decentralized networks are trying to adapt and address their challenges with bad actors.

The Future Centralized Social Media

In December 2019, Twitter CEO Jack Dorsey took to the platform to announce an initiative to create an "open and decentralized standard for social media" (Dorsey 2019). Citing a range of issues plaguing social media companies, Dorsey proclaimed that the initiative, dubbed "bluesky," would explore a standard for communication that could be used by many social media companies, including Twitter.

In his thread, Dorsey cited Mike Masnick's article, "Protocols, Not Platforms," which lays out an argument in favor of a decentralized social network (Masnick 2019). Rather than have all social interactions take place within a handful of services, a vast collection of smaller social networks could adopt open protocols for communication, much the same way that email service providers

operate using the shared, open protocols. According to Dorsey, not only would such a system encourage innovation by allowing other companies to compete with big social media providers for their users, it would improve moderation by distributing it from one central authority to a vast array of moderators.

Dorsey promised to "either find an existing decentralized standard they can help move forward, or failing that, create one from scratch." He signaled his expectation to build a new protocol from the ground up, saying "we'd expect this team not only to develop a decentralized standard for social media, but to also build open community around it."

I asked Fediverse moderators what they thought about Twitter developing an open protocol, potentially even adopting the ActivityPub standard upon which their own networks depended. Reactions ranged from concern to indifference. "I don't think they will implement ActivityPub," said Catgirl. "The amount of tweets would flood other instances. Many instance owners told me they will block Twitter." When I asked if he was worried about Twitter adopting an open protocol, Ash Furrow said, "I'll believe it when I see it."

The Future for the Fediverse

I tracked Laurelai down and asked her how she felt about the way that Mastodon and the broader Fediverse had adapted to bad actors. I was disappointed to learn that she wasn't happy: she felt that Mastodon had failed to take some steps that would make users safer (making federation between servers a mutual act, changing the way that block rules function). After Gab had first arrived, she packed up and migrated to a new server and installed custom modifications to improve the safety of her users. She doesn't think the Fediverse will exist, in the long run. And yet, it means something that Laurelai is still here, occupying a corner of the decentralized social network. Decentralization and open protocols make it possible for people to build communities according to the rules they want, and human moderators—context aware, embedded in the communities they serve—make these spaces safer and more reflective of their users values.

As the world grows increasingly connected and more people come online, big social media companies are vying to be the platform that hosts online human interaction. Due to their sprawling size and slowness to address fundamental content issues, these platforms are ill-equipped for creating safe spaces for people to interact. Perhaps, as decentralized networks grow in scale and sophistication, they will emerge as a viable alternative to the centralized networks, and prove that social spaces can be safe, responsive, and truly our own.

References

'A Further Update on New Zealand Terrorist Attack' (2019) *Facebook*, 20 March. Available at: https://about.fb.com/news/2019/03/technical-update-on-new-zealand/ (Accessed: 24 August 2020).

‘ActivityPub Is Now a W3C Recommendation | W3C News’ (2018), 23 January. Available at: https://www.w3.org/blog/news/archives/6785 (Accessed: 22 August 2020).
Basu, N. (2019) ‘SC Lawyer Sanjay Hegde Tangles with Twitter over “Defiance” Photo from 1936 Nazi Germany’, *ThePrint*, 27 October. Available at: https://theprint.in/india/sc-lawyer-sanjay-hegde-tangles-with-twitter-over-defiance-photo-from-1936-nazi-germany/312064/ (Accessed: 22 August 2020).

Behrenshausen, B. (2013) ‘Pump.io: The Decentralized Social Network That’s Really Fun’, *Opensource.com*. Available at: https://opensource.com/life/13/7/pump-io (Accessed: 22 August 2020).

Caelin, D. (2020) ‘Derek Caelin (@Argus@mastodon.technology)’, *Mastodon for Tech Folks*. Available at: https://mastodon.technology/@Argus/104178423480881678 (Accessed: 23 August 2020).
‘Chocobozzz/PeerTube’ (n.d.) *GitHub*. Available at: https://github.com/Chocobozzz/PeerTube (Accessed: 23 August 2020).

Statista Research Department, (2021) ‘Facebook: Active Users Worldwide’, *Statista*. Available at: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/ (Accessed: 03 December 2021).

Statista Research Department, (2021) ‘Twitter: Number of Employees 2019’, *Statista*. Available at: https://www.statista.com/statistics/272140/employees-of-twitter/ (Accessed: 03 December 2021).

‘Community Standards’ (n.d.) *Facebook*. Available at: https://www.facebook.com/communitystandards/objectionable_content (Accessed: 21 August 2020).

Conway, M., Scrivens, R. and Macnair, L. (2019) ‘Right-Wing Extremists’ Persistent Online Presence: History and Contemporary Trends’, *International Centre for Counter-Terrorism*. Available at: https://www.jstor.org/stable/resrep19623 (Accessed: 17 September 2020).

Copia Institute (2020) ‘Content Moderation Case Study: Facebook’s AI Continues to Struggle with Identifying Nudity’. Available at: https://www.techdirt.com/articles/20201211/15111245871/content-moderation-case-study-facebooks-ai-continues-to-struggle-with-identifying-nudity-2020.shtml (Accessed: 31 August 2021).

Cox, J. and Kobler, J. (2019) ‘Facebook Bans White Nationalism and White Separatism, Motherboard’. Available at: https://www.vice.com/en_us/article/nexpbx/facebook-bans-white-nationalism-and-white-separatism (Accessed: 22 August 2020).

‘Dangerous Speech: A Practical Guide’ (2020) *Dangerous Speech Project*. Available at: https://dangerousspeech.org/guide/ (Accessed: 1 October 2020).

Dorsey, J. (2019) ‘Jack on Twitter’, *Twitter*. Available at: https://twitter.com/jack/status/1204766078468911106 (Accessed: 22 August 2020).

European Commission (2020) ‘The Code of Conduct on Countering Illegal Hate Speech Online’, *European Commission*. Available at: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_1135 (Accessed: 24 August 2020).

‘Facebook World Stats and Penetration in the World—Facebook Statistics’ (2020) *Internet World Stats*. Available at: https://internetworldstats.com/facebook.htm (Accessed: 1 October 2020).

Framasoft, ‘V3 Roadmap Moderation, Let’s Improve Plugins and Playlists!’ (2020) *JoinPeerTube*. Available at: https://joinpeertube.org (Accessed: 23 August 2020).
Gans, J. and Leigh, A. (2019) *Innovation + Equality: How to Create a Future That Is More Star Trek Than Terminator*. Cambridge: MIT Press.

Gupta, Anika (2020) *How to Handle a Crowd*. Tiller Press
‘Hateful Conduct Policy’ (n.d.) Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (Accessed: 22 August 2020).

IsolateGab (2020) ‘IsolateGab: mastodon: (@isolategab@todon.nl)’, *Todon.nl*. Available at: https://todon.nl/@isolategab/104254977308985846 (Accessed: 24 August 2020).

Jackson, Gita (2018) ‘Twitter’s Emerging Competitor Doesn’t Want To Be Twitter’, *Kotaku*. Available at: https://kotaku.com/twitters-emerging-competitor-doesnt-want-to-be-twitter-1829069122 (Accessed: 1 October 2020).

Katt, fracturing (2020) ‘katt, fracturing (@starkatt@vulpine.club)’, *The Vulpine Club*. Available at: https://vulpine.club/@starkatt/104210054305766763 (Accessed: 23 August 2020).

Khan, F. (2019) ‘Twitter Won’t Restore Sanjay Hegde’s Account, Lawyer Says Will Move Court’, *ThePrint*, 5 November. Available at: https://theprint.in/india/twitter-wont-restore-sanjay-hegdes-account-lawyer-says-will-move-court/315904/ (Accessed: 22 August 2020).

Klonic, K (2017) ‘The New Governors: The People, Rules, and Processes Governing Online Speech’. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937985 (Accessed: 31 August 2021).

Laurelai (2019) ‘Hardcode a Block for Gab’s Instance by Default · Issue #108 · Florence-Social/Mastodon-Fork’, *GitHub*. Available at: https://github.com/florence-social/mastodon-fork/issues/108 (Accessed: 22 August 2020).

Masnick, M. (2019a) ‘Masnick’s Impossibility Theorem: Content Moderation At Scale Is Impossible to Do Well’, *Techdirt.* Available at: https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml (Accessed: 22 August 2020).

Masnick, M. (2019b) ‘Protocols, Not Platforms: A Technological Approach to Free Speech, Knight First Amendment Institute at Columbia University’. Available at: https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech (Accessed: 22 August 2020).

Morse, J. (2017) ‘Bye, Twitter. All the Cool Kids Are Migrating to Mastodon’, *Mashable*. Available at: https://mashable.com/2017/04/04/mastodon-twitter-social-network/ (Accessed: 1 October 2020).

Murphy, Laura and Cacace, Megan (2020) ‘Facebook’s Civil Rights Audit—Final Report’, *Facebook*, p. 42. Available at: https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf (Accessed: 30 September 2020).

‘Netizens Accuse Twitter of Caste Bias for Verifying Users in India’ (2019) *News18*. Available at: https://www.news18.com/news/tech/netizens-accuse-twitter-of-caste-bias-for-verifying-users-in-india-2376411.html (Accessed: 30 September 2020).

Octodon (n.d.) ‘Mastodon Hosted on octodon.social’. Available at: https://octodon.social/about/more (Accessed: 23 August 2020).

O’Donovan, C. et al. (2019) ‘YouTube Is Still Struggling to Rein in Its Recommendation Algorithm’. Available at: https://www.buzzfeednews.com/article/carolineodonovan/down-youtubes-recommendation-rabbithole (Accessed: 23 August 2020).

PeaceTech Lab | Hate Speech (n.d.) ‘PeaceTech Lab | Putting the Right Tools in the Right Hands to Build Peace’. Available at: https://www.peacetechlab.org/hate-speech (Accessed: 23 August 2020).

Puffball (2020) *puffball (@wgahnagl@www.librepunk.club)*, *librepunk* Available at: https://www.librepunk.club/@wgahnagl/104210345911497659 (Accessed: 23 August 2020).

Qoto Mastodon (n.d.) ‘Mastodon hosted on qoto.org’. Available at: https://qoto.org/about/more (Accessed: 22 August 2020).

Rhodes, M. (2017) ‘Like Twitter But Hate the Trolls? Try Mastodon’, *Wired*, 13 April. Available at: https://www.wired.com/2017/04/like-twitter-hate-trolls-try-mastodon/ (Accessed: 1 October 2020).

Robertson, A. (2019) ‘How the Biggest Decentralized Social Network Is Dealing with Its Nazi Problem’, *The Verge*. Available at: https://www.theverge.com/2019/7/12/20691957/mastodon-decentralized-social-network-gab-migration-fediverse-app-blocking (Accessed: 22 August 2020).

Rochko, E. (2017) ‘Learning from Twitter’s Mistakes’, *Official Mastodon Blog*. Available at: http://blog.joinmastodon.org/2017/03/learning-from-twitters-mistakes/ (Accessed: 22 August 2020).

Rules | Mastodon.technology (n.d.) ‘Mastodon Hosted on mastodon.technology’. Available at: https://mastodon.technology/about/more (Accessed: 22 August 2020).

Rosenberg, A. (2018) ‘Gab, a Racist-Friendly Alt-Twitter, Has Been Banned by PayPal and Others’, *Mashable*. Available at: https://mashable.com/article/gab-paypal-stripe-joyent-synagogue-shooting/ (Accessed: 1 October 2020).

Sample, I. (2019) ‘Study Blames YouTube for Rise in Number of Flat Earthers’. *Science | The Guardian*. Available at: https://www.theguardian.com/science/2019/feb/17/study-blames-youtube-for-rise-in-number-of-flat-earthers (Accessed: 23 August 2020).
Schulze, E. (2019) ‘EU Says Facebook, Google and Twitter are Getting Faster at Removing Hate Speech Online’, *CNBC*. Available at: https://www.cnbc.com/2019/02/04/facebook-google-and-twitter-are-getting-faster-at-removing-hate-speech-online-eu-finds--.html (Accessed: 22 August 2020).

‘Telecommunications Act of 1996 | United States [1996]’ (n.d.) *Encyclopedia Britannica*. Available at: https://www.britannica.com/topic/Telecommunications-Act (Accessed: 23 August 2020).

‘The Commons, an Intervention to Depolarize Political Conversations on Twitter and Facebook in the USA, 2019 Report’ (n.d.). *Build Up*, p. 3. Available at: https://howtobuildup.org/wp-content/uploads/2020/04/TheCommons-2019-Report_final.pdf (Accessed: 30 September 2020).
Tufekci, Zeynep (2017) *Twitter and Teargas*. Yale University Press

Vaidhyanathan, Siva (2018) *Antisocial Media: How FAcebook Disconnects Us and Undermines Democracy*. Oxford University Press.

Valens, A. (2019) ‘Mastodon Is Crumbling—And Many Blame Its Creator’, *The Daily Dot*. Available at: https://www.dailydot.com/debug/mastodon-fediverse-eugen-rochko/ (Accessed: 23 August 2020).

Van Dijck, J., Poell, T. and De Waal, M. (2019) *The Platform Society: Public Values in A Connective World*. New York: Oxford University Press.

Wong, Q. (2019) ‘Facebook Content Moderation Is an Ugly Business. Here’s Who Does It’, *CNET*. Available at: https://www.cnet.com/news/facebook-content-moderation-is-an-ugly-business-heres-who-does-it/ (Accessed: 15 October 2020).

Wright, L. (2019) ‘Twitter Bans Religious Dehumanization’, *Dangerous Speech Project*. Available at: https://dangerousspeech.org/twitter-bans-religious-dehumanization/ (Accessed: 22 August 2020).

Yowlen (2019) ‘Fedilab, Librem Social Supports the Pushing Out of Marginalized Groups - Apps - F-Droid Forum’, *F-Droid Forums*. Available at: https://forum.f-droid.org/t/fedilab-librem-social-supports-the-pushing-out-of-marginalized-groups/6470 (Accessed: 24 August 2020).